THE NLP PIPELINE FOR LEGAL TEXT PROCESSING

By Aryan Deshpande - 250215

SENTENCE SEGMENTATION

- Segmentation refers to the process of dividing text into smaller, meaningful units such as sentences or words. This is a foundational step in many NLP tasks, as it structures raw text into manageable components for further analysis.

- Sentence segmentation, also known as sentence boundary detection, involves splitting a body of text into individual sentences. This step is critical for tasks like parsing, machine translation, and summarization, as these algorithms rely on well-defined sentence boundaries for accurate processing.

TOKENZATION

- Once we split our document into sentences, we need to break this sentence into separate words or tokens. This is called tokenization.
- We'll just split apart words whenever there's a space between them.
- we'll also treat punctuation marks as separate tokens since punctuation also has meaning.

STOP-WORDS REMOVAL

- Stop words are those words that you might want to filter out before doing any statistical analysis.
- For example, the stop words could be “and”, “the”, and “a”. These words are usually around the key words enhancing grammar.
- There’s no standard list of stop words that is appropriate for all applications. The list of words to ignore can vary depending on your application.

LEMMATIZATION

- Lemmatization is a fundamental text preprocessing technique in Natural Language Processing (NLP).
- It is the process of figuring out the most basic form or lemma of each word in the sentence.
- Lemmatization can be divided into different categories based on the word type and context. It could be based on verb, noun or adjective.

PART-OF-SPEECH (POS) TAGGING

- It is a fundamental task where each word in a sentence is assigned a grammatical category such as noun, verb, adjective or adverb.
- Knowing the role of each word in the sentence will help us start to figure out what the sentence is talking about.
- It is essential for understanding the structure and meaning of the text data.

NAMED ENTITY RECOGNITION (NER)

- NER is to detect and label the nouns with the real-world concepts that they represent.
- It identifies and classifies key elements in text—called entities—into predefined categories such as Person, Organization, Location, Date, Quantity, and more. It transforms unstructured text into structured data.
- NER systems use the context of how a word appears in the sentence and a statistical model to guess which type of noun a word represents.
- A good NER system can tell the difference between "*Brooklyn Decker*" the person and the place "*Brooklyn*" using context clues.